DOCUMENT RESUME

ED 287 283 FL 016 939

AUTHOR Larson, Jerry W.
TITLE Computerized Adaptive Language Testing: A Spanish Placement Exam.
PUB DATE 87
NOTE 11p.; In: Language Testing Research; Selected Papers from the Colloquium (Monterey, California, February 27-28, 1986); see FL 016 938.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; Higher Education; *Language Tests; Latent Trait Theory; Second Language Instruction; *Spanish; *Student Placement; Test Construction; Test Reliability
IDENTIFIERS *Brigham Young University UT

ABSTRACT

The development of a Spanish computerized adaptive placement test at Brigham Young University (Utah) is described. The test is based on an approach that locates an examinee's ability along a continuum of latent ability rather than determining the degree to which an individual's ability differs from the abilities of others taking the test, as in conventional tests. The report gives background information about the principles of adaptive testing, the rationales of test design, the use of computers for adaptive testing, and the status of computerized adaptive language testing, and outlines the university's procedure for developing a specific test for Spanish. That process includes the following: (1) development of an item bank; (2) determination of a suitable psychometric methodology; (3) selection and development of an appropriate test delivery system; and (4) test validation and implementation. The test that has been developed shows great promise. Further research on the application of computerized adaptive testing for proficiency, achievement, and diagnosis, and the expansion of the testing technology to include a listening component are recommended. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. Schaffman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED287283

COMPUTERIZED ADAPTIVE LANGUAGE TESTING: A SPANISH PLACEMENT EXAM

Jerry W. Larson
Brigham Young University

Language Placement Testing

Placement testing is a concern of nearly all foreign language departments in institutions of higher education. Each semester in-coming students must be counseled and placed into courses that are best suited to their needs and abilities; they are often directed to particular courses based solely on the number of classes they have taken in secondary school, with no formal placement measure being used. This placement procedure does not take into account several factors, such as the effectiveness of previous teachers, the specific information covered, extracurricular or other out-of-class exposure to the language, and, in general, how well the student can actually use the language.

Among the reasons for this method of placement may be a lack of efficient and accurate placement tests, or insufficient time required for testing and evaluating substantial numbers of students at the beginning of the semester. Testing each student before placement is difficult, since adequate placement tests traditionally have required a significant amount of time--about two to two and a half hours for a fairly thorough placement exam--and a lot of coordination effort, i.e., arranging testing times and locations, administering tests, scoring answer sheets, recording and reporting the results, and finally counseling the students.

A search of the literature yields comparatively little discussion about the development and use of adequate foreign language placement measures; nevertheless, conversations with colleagues indicate there is considerable concern about this area of testing and the paucity of research and development being done. Given the diversity of student backgrounds, learning styles, and levels of preparation, placement tests are desperately needed that are able to "predict success in one course and discourage or reject placement in another" (Callas, 1981, p.27).

Adaptive Testing

Adaptive testing is based upon item response theory, a statistical theory which assumes that there is an underlying latent ability distribution for any ability. The ability level of each individual taking a given test lies somewhere along the latent ability continuum. An adaptive test is designed to locate an examinee's ability along that continuum, rather than to determine the degree to which an individual's ability differs from the abilities of others taking the same measure, as is the case with conventional tests (Olsen, 1982, p.1). Traditional tests present a common set of items for all students taking the test, while adaptive tests tailor the test item difficulty to match the ability level of the individual.

FL016939

The concept of adaptive testing has existed for several years. The early version of the Stanford-Binet intelligence test, for example, was an adaptive test. It was administered by a trained examiner who, during the test, selected subsequent test items according to the individual examinee's responses to previous questions. However, when it became necessary to test large numbers of students, the test became less adaptive in nature.

Testing Rationales

Test design has generally been guided by one of two rationales. One is to devise a "peak test," a measure that contains items fairly homogeneous in difficulty and that is able to discriminate quite effectively over a very narrow range of the variable being tested but that discriminates poorly outside that range. The other rationale is to develop a "uniform test," one comprising items that are heterogeneous in difficulty and thus test a wide range of ability but with a much lower level of precision.

A third rationale, however, is the "adaptive test," which "can approximately achieve the high point precision of a peaked test and can extend that high level of precision over the wide range of a uniform test" (McBride, 1979, p.3). From a psychometric viewpoint, adaptive tests have the additional advantage of a small measurement error, since they concentrate on a single ability level, yet they test a wide range of the trait scale. They are also able to yield precise measurement with fewer test items for the candidate to answer and with increased reliability and external validity (McBride, 1979, pp.1-4).

Adaptive Tests and Computers

Adaptive tests can be administered in three ways: 1) by a skilled examiner, 2) using a specially designed paper-and-pencil test, or 3) at a computer terminal. If only a few students are required to take the test, it may be possible to have skilled testers who can administer the test to each person individually. But, if large numbers of students are to be tested, using trained examiners in this way is impractical. Paper-and-pencil type adaptive tests are possible with either small or large groups, although certain disadvantages must be considered. In order for a paper-and-pencil test to function as an adaptive test, it would require extensive, complex branching instructions for the examinee to follow. This may confuse the dimension to be measured with the examinee's ability to follow instructions (McBride, 1979, p.24). Pencil-and-paper adaptive tests are also subject to the problems of test security, reprinting, scoring and reporting results.

Among the benefits peculiar to computerized adaptive tests are the following:

a. Affective benefits to examinees:

(1) Since this type of test is tailored to the individual, the examinee does not have to experience the boredom or frustration commonly associated with taking lengthy tests, which often range in difficulty from extremely simple to very difficult.

(2) The examinee spends less time taking the test, since fewer items will be necessary for evaluating performance levels.

(3) Examinees receive immediate feedback, without having to wait for days for their results.

(4) Examinees are able to work at their own pace.

(5) Computerized adaptive testing relieves the student of having to comply with complex testing instructions, since the computer automatically selects the appropriate branching routines.

b. Administrative benefits:

(1) Because the computer controls the branching and individualization of test item sequencing, fewer test administrators are needed.

(2) Less time is needed to administer each individualized test.

(3) Test security is improved.

(4) In the event that some test items need changing, a simple deletion and addition can be accomplished without having to reprint new test booklets.

(5) Pretesting of potential test items is easily accomplished without having them recorded as part of the examinee's score.

(6) Computerized adaptive testing eliminates the problem of cheating and test compromise, since each test is virtually unique.

(7) The headaches of test storage, handling test booklets and answer sheets (and their costs) are eliminated.

(8) Also alleviated are the time and expense involved in scoring and reporting test results.

Computer Adaptive Language Testing (CALT)

Although computer adaptive testing has existed in some fields during the past few years, it is a relatively new to foreign language testing. CALT appears to be very well suited to various kinds of foreign language testing purposes, e.g., placement, diagnostic, and proficiency testing, yet it has been slow finding its place in foreign language testing programs. Considering the increased accessability and use of computers in language instruction, it appears that the time is right for serious efforts toward developing computerized adaptive foreign language tests.

Seminal development of CALT placement and proficiency tests is starting to surface at universities, testing services, and other organizations involved with foreign language teaching and testing. Researchers at Brigham Young University have recently developed CALT placement tests for ESL and Spanish; language testing experts at the University of California, Los Angeles, have

produced a CALT test for ESL; the College Board has announced a new computerized placement test for Freshman English programs; and ATFL is supporting development of a CALT proficiency test.

Development of a Spanish Computerized Adaptive Placement Exam

The Department of Spanish and Portuguese at Brigham Young University is faced with placing hundreds of in-coming students in lower division courses each year. Because of the enormity of the task, and considering the time and manpower that would be required, it seemed judicious to utilize the computer to the extent possible to assist in this effort.

Research was undertaken to develop a Spanish computerized adaptive placement exam (S-CAPE). Not only was the computer to serve as the "administrator" of the test but also as the manager, or record keeper, of the test results. Additionally, in its role as test administrator, the computer was to adapt the test to the individual examinee, i.e., present items whose difficulty level matched the ability level of the examinee.

Four principal steps were followed in producing this CALT test: 1) development of a test item bank, 2) determination of a suitable psychometric methodology, 3) selection and development of an appropriate S-CAPE delivery system, and 4) validation and implementation of the test.

Development of the Test Item Bank

A critical step in the development of an effective computerized adaptive placement test is having a substantial number of accurately calibrated test items in the item bank. Since the bulk of the students entering Spanish classes must be placed primarily in one of the first four levels, there must be a sufficient number of items at each of those levels to provide accurate indications of student ability. The task of writing test items to cover the four levels of instruction (Spanish 101, 102, 201 and 202) began in October, 1984. Several competent upper-division and graduate students were recommended by the Department of Spanish and Portuguese to work on the project. Each person was interviewed to determine whether his or her background and experience were suited for the task. Then four research assistants, two native Spanish-speakers and two non-natives, were chosen to help with the project.

The project required items to test grammar, reading, and vocabulary for each of the four levels. 1,443 items were written and given to members of the Spanish faculty, who were asked to judge the level and quality of the questions to the best of their knowledge and ability. (Criteria for "quality" included appropriateness of expression, no misleading distractors, use of authentic language, etc.). Eleven hundred of those items were selected and put into five separate test forms (A-E), each having two parts; these tests were administered to 199 students from the four lower division Spanish courses. The tests were submitted to two types of statistical analyses after students had completed them: conventional and Rasch analyses. The conventional analysis consisted of determining the Hoyt and KR-21 reliability coefficients and the point biserial (PBIS) discrimination coefficients. (PBIS is an item level correlation: right versus

wrong on an item correlated with the individual's total score.) The results are presented in Table 1.

Table 1

Reliability and Point Biserial Coefficients
of Test Forms A - E

Test Form	Part	Hoyt	KR-21	PBIS
A	1	.95	.93	.79
	2	.94	.92	.82
B	1	.96	.95	.74
	2	.95	.94	.82
C	1	.97	.96	.82
	2	.97	.96	.82
D	1	.96	.95	.79
	2	.95	.94	.83
E	1	.93	.91	.77
	2	.87	.83	.85

n = 199

The Rasch analysis was run using Microscale plus TM, a statistical program written specifically for computing Rasch fit statistics.[1] Using this program, item difficulty indices were obtained for each of the items on each of the five test forms. All of the items were "anchored" together and located along a single difficulty/ability continuum. This enabled us to select items of specified difficulty, which is crucial to constructing an adaptive test. The range of difficulty of the items was from -3.6 to $+6.0$ logits[2], which represents 96 difficulty levels. The 96 levels were ultimately condensed into 51 levels (0-50), covering a difficulty range of -2.0 to $+5.8$. This range of difficulty easily spanned the ability level of students in the lower division Spanish courses.

After the test items had been calibrated and assigned a difficulty index, they were again given to members of the Spanish Department for careful scrutiny, checking for errors in spelling, inaccuracies in grammar, misleading or tricky distractors, or any other potential problems present in the items. As a

[1] For a discussion of the appropriateness of using the Rasch model for calibrating subtests of various language skills, see Henning, Hudson, and Turner in Language Testing, 2(2) (December, 1985): 141-154.

[2] A logit is a unit of measure along the difficulty continuum.

result of this review, an additional 51 items were removed from the item bank, leaving about 1,050 items that could be used in the final placement test.

Each of the remaining items were coded, indicating item type (i.e., grammar, reading, or vocabulary), difficulty index, question number, and the letter of the correct response.[3] Nine items at each of the 51 difficulty levels were included in the final version of the placement test. These test items were placed in a random access file, which was created to provide quick and random selection of items within individual difficulty levels.

Testing Methodology

The test was designed so that the computer would present to each individual examinee items that are in the estimated range of his or her ability. Placement into a specified course would be determined by the student's performance on these items. In order to avoid a student's being given the same item twice, a sequential file was created that contained a key for each question in the test item base. This file was read into an array in memory at the beginning of each individual test session. That array was flagged to indicate whether a question had been used. If the question had appeared previously during the test in progress, it would not be selected a second time.

In order for the test to "zero in" on the ability range of an examinee, the following procedure was used: When a student begins the test, the computer selects an item from difficulty level 20, which is toward the lower end of the ability range of most beginning students. If the examinee answers the item correctly, an item at level 26 is randomly selected and displayed on the screen; if the examinee answers the initial item incorrectly, the next item that appears will be from level 15. For the first six questions, the test will increment the difficulty level by 6 for the next question if the response is correct, or it will decrease by 5 if the response is incorrect. This allows the test to "probe," or advance quickly to the approximate ability level of the examinee. After the first six questions have been attempted, the difficulty level adjusts by plus or minus one, depending on the correctness of the previous response, which allows for pinpointing the examinee's actual level of performance. The test is terminated when one of the following conditions is met: 1) the examinee has five correct responses (not necessarily in succession) at any one difficulty level, or 2) the examinee has four incorrect answers on any one level. It was found that condition #2 was the most common cause for the test terminating, since students met condition #1 only if they were answering at the highest level of difficulty (Level 50). At the conclusion of the test, the computer immediately informs the examinees of their ability level. Using this information, the student refers to a Placement Chart, which indicates the ability levels suggested for each course. These parameters, or cutoff points, were determined from the performance of the students from the courses in question, who originally served as the norming group for the test,

---

[3] The items used in this placement test are all objectively-scored multiple-choice items, focusing on either grammar analysis, reading comprehension, or vocabulary recognition.

and of the students in two validation groups (see below). Students who place near the top of a particular range of course scores are encouraged to enroll in the next higher course, challenging themselves a little more than normally would be required.

S-CAPE Delivery System

The S-CAPE (Spanish Computer Adaptive Placement Exam) is designed for administration on a standard IBM-PC, or compatible microcomputer, with 64K memory. Either a monochrome or color monitor may be used. The S-CAPE system requires either two floppy disk drives or one hard disk. With the two-drive system, drive A houses the program disk, and drive B contains the examinee's test performance data. After the test master disk is booted, instructions are displayed, informing the examinees how to proceed with the test[4] followed by instructions to enter their name, social security number, and information about their previous exposure to Spanish; this information is recorded onto each examinee's record file. After the student enters this information, the computer automatically displays the first item of the test. When the examinee has answered the appropriate number of items at his or her level of ability, the test will automatically terminate, and the student's ability level is displayed on the screen.

At the conclusion of an individual test or at the end of the day, the person in charge of the testing operation can display on the screen or print on paper all or selected student records. The floppy data disks will accommodate up to 200 different student data files. (Examinees are not made aware of the procedure for printing the test data files.) The printout of each record contains the examinee's name, social security number, amount of previous Spanish instruction or experience, date the test was taken, time the test began, time the test was completed, total number of correct and incorrect responses, and overall performance level. The printout also displays the response pattern of the examinee, indicating which items (by item identification number) were presented to the examinee, the difficulty index of each item, and whether the item was answered correctly or incorrectly. (This information may prove useful for item response analysis or for isolating particular learning deficiencies of students.) An actual sample printout of one of the students--whose name has been changed--is included in Appendix A.

Implementation of the S-CAPE System

Validation studies were conducted before using the S-CAPE on a large-scale basis. The validation sessions were conducted using students from the courses for which the test was designed. Seventy-four students (ranging in ability from only a brief encounter in high school to students completing Spanish

---

[4] The instructions to the student simply explain that they are to respond by typing the letter of the correct answer choice. All keyboard keys other than those corresponding to answer choices are disabled, thus preventing any disconcerting error messages or other distractions from the task at hand.

302) took the test during the first validation session. This was done during the final week of Fall Semester, 1985, when selected teachers sent to take the placement test students whom they judged definitely ready to move on to the next course level. It was believed that the responses and level designations of these students would provide accurate information for determining the course level cutoff points. However, some students who were not ready to advance to the next higher course inadvertently took the test as well; therefore, it was necessary to review the outcome of the testing session on a student-by-student basis before being able to determine the minimum ability level required for each course.

A second validation session was conducted during Winter Semester, 1986, using 129 students.

Cutoff levels have now been determined specifically for the Department of Spanish and Portuguese at Brigham Young University, which is using the S-CAPE to screen and place students into its courses. Using fifteen microcomputers in the Humanities Learning Resource Center at Brigham Young University, it is possible to test 300-350 students in an eight-hour day, since the tests require on the average approximately 18-20 minutes to complete.[5]

Conclusions and Recommendations

While there is no such thing as a perfect placement test, initial observations and evaluations of the S-CAPE are extremely positive and promising. Further research is recommended, however, in order to produce a test in which one would have near total confidence. For example, additional research might be done to perfect the procedure for determining test termination criteria.

Further research into and application of the computerized adaptive testing format should occur in the areas of proficiency, achievement, and diagnostic testing.

Given the technology of the day, it is possible to expand the scope of computerized adaptive language testing by adding a listening component, using a random-access audio disk player to select listening items that have been properly calibrated. Audio disk technology allows for almost instantaneous retrieval of any prerecorded segments on the disk. The computer could be programmed to seek and retrieve audio segments based upon examinee responses.

The results of the efforts to develop a Spanish computerized adaptive placement test are gratifying. This project, as well as others mentioned above, have shown that it is possible to produce tests that can substantially reduce the burden of language placement testing for language departments with large enrollments.

---

[5] The minimum amount of time spent by an examinee taking the test is six minutes, while the maximum has been 59 minutes.

References

Callas, Dennis. 1981. Assessment data, the academic system and evaluation of the placement process. Community College Review, 9(2), 27-33.

Henning, Grant, Hudson, Thom, and Turner, Jean. 1985. Item response theory and the assumption of unidimensionality for language test. Language Testing, 2(2), 141-154.

McBride, James R. 1979. Adaptive mental testing: the state of the art. Technical Report 423. Alexandria, Virginia: U.S. Army Research Institute for the Behavioral and Social Sciences.

Olsen, James B. 1982. Psychometric recommendations for computerized adaptive testing. Monograph. Orem, Utah: WICAT Systems.

Appendix A

Spanish Computer Adaptive Placement Exam Report

Jane Jones 576-56-0823 Date: 04-02-1986 Time started: 11:56:06

Spanish Background:
One semester in college

Difficulty Level	Test Item	Student Answer	Right/ Wrong
20	174	A	w
15	123	D	r
21	182	C	w
16	134	B	r
22	191	B	r
28	244	A	w
23	202	C	w
22	188	B	r
23	197	A	r
24	212	D	r
25	220	D	w
24	206	B	w
23	198	B	r
24	205	C	r
25	218	A	w
24	210	A	w
23	204	D	r
24	213	B	w
23	196	C	r
24	211	A	w

There were four incorrect answers at level 24
LEVEL: 24 Number right: 23

Test completed at 12:07:52